PLANNING AND DESIGN OF SURVEYS

2-1 Design of Surveys

2-2 Survey Response Rate Parameters

2-3 Developing a Request for Proposals (RFP) for Surveys

2-4 Pretesting Survey Systems

2-5 Maintaining Data Series over Time

2-6 Educational Testing

SUBJECT: DESIGN OF SURVEYS

NCES STANDARD: 2-1

PURPOSE: To identify the survey design components required to conduct a data collection.

KEY TERMS: confidentiality, domain, effective sample size, estimation, field test, frame, individually identifiable data, key variables, planning document, precision, probability of selection, response rate, strata, survey, survey system, target population, and variance.

STANDARD 2-1-1: A technical document that delineates the basic design of a survey or survey system must be developed prior to the initiation of a data collection. The document must address the objectives of the survey as indicated in the initial planning document; the survey design; the data collection plan; and the personnel resources, funds, and time needed to achieve high data quality. To meet this standard, the survey design plan must include the following:

1. A detailed discussion of the goals and objectives of the survey or survey system, including the information needs that will be met, content areas included, target population(s), and analytic goals.
2. A discussion of the sample design that describes how it will yield the data required to meet the objectives of the survey. The discussion must identify the following:
   - Sampling frame and the adequacy of the frame (see Standard 3-1);
   - Sampling unit for each stage of data collection;
   - Criteria for stratifying or clustering;
   - Sampling strata;
   - Minimum substantively significant effect (e.g., the smallest differences the survey is intended to measure) for the total population;
   - Minimum substantively significant effect for key reporting variables by reporting domains;
   - Power analyses to determine sample sizes;
   - Effective sample sizes for key variables by reporting domains;
   - Sample size by stratum;
   - Known probability of selection;
   - Expected yield by stratum;
   - Estimated efficiency of sample design;
   - Weighting plan;
   - Variance estimation techniques appropriate to the survey design; and
   - Expected precision of estimates for key variables.
3. A listing of all survey data items, including time series data items, how each item can best be measured (e.g., through questionnaires, assessments), and reasonable evidence that these items are valid and can be measured both accurately and reliably.

4. An analysis plan providing evidence that the basic information needs which justify the study can be met through the proposed data collection. The data element plan must demonstrate how the proposed sample, the survey items, and the measurement methods are related to the objectives of the survey. The data element plan must identify analysis issues, key variables, minimum substantively significant effect sizes, and proposed statistical tests (see Standard 5-1).
5. The anticipated data collection procedures, including the following:
   - Frequency and timing of data collection;
   - Primary mode of collection;
   - Data collection protocol for field staff, including refresher training on a routine, recurring cycle;
   - Training of survey collection staff and persons coding and editing the data;
   - Expected unit response rates for each stage of data collection (see Standard 2-2); and
   - Methods for achieving acceptable response rates (see Standard 3-2).
6. A nondisclosure pledge (see Standard 4-2-4).
7. A security plan for preserving the confidentiality of the data during collection, processing, and analysis, if individually identifiable data will be collected (see Standard 4-2-5).
8. A disclosure analysis plan that describes how disclosure risk will be controlled through the use of data perturbation and coarsening associated with confidentiality edits (see Standard 4-2-7 and Standard 4-2-8).
9. A plan for producing and disseminating all relevant data files for data release (e.g., public use, restricted use, on-line analysis tool) (see Standard 7-1).
10. An outline of a plan for quality assurance during each phase of the survey process that will permit monitoring and assessing the performance during implementation. The plan must include contingencies to modify the survey procedures, if design parameters appear unlikely to meet expectations (for example, low response rates) (see Standard 3-3).
11. A plan for pretesting items and administration protocols for the survey or survey system (this includes both pilot and field tests) (see Standard 2-4).
12. An outline of the general parameters for evaluating survey procedures and results (see Standard 4-3).
13. General specifications for an internal project management system that identifies critical activities and key milestones of the survey that will be monitored, and the time relationships among them (see Standard 3-3).
14. An estimate of the target time period needed for the full survey cycle, including the estimated times for the following:
   - Planning and development (including pretesting);
   - Data collection;
   - Processing and data editing;
   - Weighting;
   - Imputations, if needed;
   - Disclosure avoidance plan and analysis;

- File construction;
- Survey documentation; and
- Preparation and review of the initial release report.

The time from end of data collection to the release of the initial report should not exceed one year.

15. An Independent Government Cost Estimate (IGCE) for the entire study, including the each of the items listed in 14.

SUBJECT: SURVEY RESPONSE RATE PARAMETERS

NCES STANDARD: 2-2

PURPOSE: To specify design parameters for survey response rates. High survey response rates help to ensure that survey results are representative of the target population. Surveys conducted by or for NCES must be designed and executed to meet the highest practical rates of response. To ensure that nonresponse bias analyses are conducted when response rates suggest the potential for bias to occur.

KEY TERMS: adaptive design, assessment, cross-sectional, key variables, longitudinal, nonresponse bias, response rate, responsive design, stage of data collection, substitution, survey, target population, and universe.

STANDARD 2-2-1: Universe data collections must be designed to meet a target unit response rate of at least 95 percent.

> **GUIDELINE 2-2-1A:** A unit-level nonresponse bias analysis is *recommended* in the case where the universe survey unit response rate is less than 90 percent. (See Standard 4-4 for a discussion of nonresponse bias analysis.)

STANDARD 2-2-2: Sample survey unit response rates must be calculated without substitutions (see Standard 1-3).). To maximize planning to realistic attainable target response rates, NCES sample survey data collections must be designed to meet unit-level response rate parameters that meet or exceed historical response rates from surveys conducted with best practices and/or design targets that are deemed minimally acceptable. In addition, consideration should be given to the use of an adaptive or responsive design to increase the likelihood of obtaining a representative sample. (See Guideline 3-2-1D for a discussion of the use of response propensities in an adaptive/responsive design).

> **GUIDELINE 2-2-2A:** The following parameters represent a balance between recent NCES historical experiences and in some cases design targets that are deemed minimally acceptable:
>
> 1. For longitudinal sample surveys, the target school-level unit response rate should be at least 70 percent. In the base year and each follow-up, the target unit response rates at each additional stage should be at least 90 percent.
> 2. For cross-sectional samples, the target unit response rate should be at least 85 percent at each stage of data collection.
> 3. For random-digit dial sample surveys, the target unit response rate should be at least 70 percent for the screener and at least 90 percent for each survey component.
> 4. For household sample surveys, the target response rates should be at least 85 percent for the screener and at least 85 percent for the respondents.
> 5. For assessments, the target response rate should be at least 80 percent for schools and at least 85 percent for students.

Stage-specific design response rates, by type of survey

Type of survey	Stage-specific design response rates		
	Screener	School	All other
Universe	—	—	95
Cross-sectional	—	85	85
Longitudinal	—	70	90
Assessment	—	80	85
Random-digit dial	70	—	90
Household	85	—	85

STANDARD 2-2-3: NCES sample survey data collections must be designed to meet a target item response rate of at least 90 percent for each key item.

STANDARD 2-2-4: A nonresponse bias analysis is *required* at any stage of a data collection with a unit response rate less than 85 percent. In cases where target response rates area below 85 percent, the data collection must be designed to include nonresponse bias analyses. If the item response rate is below 85 percent for any items used in a report, a nonresponse bias analysis is also *required* for each of those items (this does not include individual test items). The extent of the analysis must reflect the magnitude of the nonresponse (see Standard 4-4).

In longitudinal sample surveys, item nonresponse bias analyses need only be done once for any individual item, unless there is a substantial deterioration in the item response rate.

STANDARD 2-2-5: In cases where prior experience suggests the potential for an *overall* unit response rate of less than 50 percent, the decision to proceed with data collection must be made in consultation with the Associate Commissioner, Chief Statistician, and Commissioner.

SUBJECT: DEVELOPING A REQUEST FOR PROPOSALS (RFP) FOR SURVEYS

NCES STANDARD: 2-3

PURPOSE: To assist NCES staff in the preparation of high quality RFPs. Each RFP should include the information required to allow an offeror to submit a proposal that demonstrates technical and managerial competence sufficient to complete successfully all phases of surveys. Each RFP should include evaluation criteria to assist the government in selecting the best offeror to conduct the work. The RFP should provide a clear, precise, and accurate description of the requirement for the work and the expected activities, services, products, and level of effort to be delivered under the contract.

KEY TERMS: survey and survey system.

STANDARD 2-3-1: RFPs must be written in compliance with guidelines established in the Federal Acquisition Regulations (FAR) and in other departmental administrative procedures and guidelines.

> **GUIDELINE 2-3-1A:** The contracting office of the Department of Education is responsible for the acquisition process for NCES and can provide expertise and guidance in the development of the RFP.
>
> **GUIDELINE 2-3-1B:** Within NCES, the staff member who is responsible for the development of a Performance Work Statement (PWS) and related documents should also be the designated Contracting Officer's Representative (COR) for the work. The staff member responsible for the development of the PWS should have completed courses required for COR certification. Minimally, the individual designated as COR should be included in the development process, to provide familiarity with the contractual requirements and expectations.

STANDARD 2-3-2: The Performance Work Statement (PWS) must contain technical, managerial, and deliverable specifications (see Standard 1-1 and Standard 2-2).

> **GUIDELINE 2-3-2A:** The technical specifications for all phases of design, implementation, and analysis include methodological, statistical, timeline, resources, analysis, and data file parameters. These specifications must be sufficiently specific to provide the information needed for a contractor to respond to a fixed price proposal. Managerial specifications should be written as specific activities and tasks. The tasks to be performed by the contractor and those to be performed by NCES should be clearly delineated. There should be a schedule for all deliverables (e.g., data files for review at various stages in processing and preparation, including the supporting computer program documentation; final data file, data file documentation, and survey methodology report; analysis plans; and final reports).

STANDARD 2-3-3: The COR must be fully certified and must maintain COR certification. COR certification requires training on contracting overview, conducting market research, independent government cost estimates, preparing statements of work, and contract administration. To maintain COR certification, the COR must adhere to training requirements established by the Department of Education's Contracts Office; this may include periodic required training on the Department of Education's financial management system, EDCAPS, and the Contracts and Purchasing Software System (CPSS).

STANDARD 2-3-4: The COR must develop an Independent Government Cost Estimate (IGCE) that includes estimates of the cost of the project for all phases and elements of the survey system or analytic work requested in terms of the contractor's manpower commitment by labor categories and other related costs. Automated Data Processing (ADP) cost, or Information Technology (IT) costs, must be estimated within each of the budget categories, to yield an estimate of total ADP costs within the total budget. Total estimated cost must not exceed the NCES budget amount for the project.

> **GUIDELINE 2-3-4A:** For further information, consult previous comparable project estimates.

STANDARD 2-3-5: To obtain funding commitment, the COR must initiate the authorization and have it approved by the Division's Associate Commissioner. The COR must confirm the survey's fiscal year scheduled activity and obtain all accounting information from the Branch Chief or the Associate Commissioner. The OC will be responsible for ensuring that the survey funds are committed in the Department's financial system and that the authorization is submitted electronically to the Contracting Officer (CO).

STANDARD 2-3-6: The Proposal Evaluation Plan specifies the members of the Technical Evaluation Panel (TEP), who serve as advisers to the Contracting Officer (CO). The plan also provides the criteria on which the COR and the TEP assess the proposals. The COR, in collaboration with the CO, assigns the factors and weights associated with each criterion. Only criteria and weights stated in the RFP may be used to evaluate submitted proposals (see Standards 1-1 and 2-2).

> **GUIDELINE 2-3-6A:** The criteria may include such factors as technical competence (e.g., qualifications of project director and staff), analysis plan, familiarity with data files, management plan, firm's organizational capacity, and past performance.

STANDARD 2-3-7: The Proposal Preparation Instructions inform the offeror as to the substantive, format, and organizational requirements for completing their proposal. The offeror must submit two separate proposals: (1) technical and (2) business. They are evaluated separately.

STANDARD 2-3-8: The COR must prepare the required clearances and approvals for the planned survey activity. The standard clearances for all new RFPs are currently the Information Technology (IT) Resources clearance, Labor Impact Determination clearance, Media Release clearance, and the Administrative Test for Characterizing Particular Services as "Personal" or "Nonpersonal" clearance.

GUIDELINE 2-3-8A: The Performance Work Statement of Request for Proposals for each survey may have its own applicable/special clearances depending on the type of procurement required. (The ACS Departmental Directive, OCFO: 2-107, Acquisition Planning, dated 2/11/2008 or later should be referenced to explain the standard clearances noted above and possible other clearances or approvals that might be required.)

STANDARD 2-3-9: The Incentive Plan, if applicable, for a performance-based contract must include a description of deliverables, schedules, and other evaluation criteria. It must also provide definitions of quality for each criterion and the associated incentive award or penalty. The evaluation criteria must include, but are not limited to, the definition of the work in measurable and/or mission-related terms.

GUIDELINE 2-3-9A: This plan tells the contractor what activity or product is required to be considered in the incentive plan. It also tells the contractor when penalties may be applied.

GUIDELINE 2-3-9B: Incentives criteria frequently include such factors as quantity, timeliness, or quality. Other criteria that are sometimes used include commercial or industry-wide standards that are used to measure performance.

GUIDELINE 2-3-9C: The following documents offer specific guidance on how to develop a performance-based solicitation:

1. Federal Acquisition Circular 97-1, https://www.acquisition.gov/far/fac/fac97-01.pdf
2. Federal Acquisition Regulation Subpart 37.6, Performance-Based Acquisition http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=675bdbc6cd84d0f3faedf5301a0615f0&rgn=div6&view=text&node=48:1.0.1.6.36.6&idno=48

SUBJECT: PRETESTING SURVEY SYSTEMS

NCES STANDARD: 2-4

PURPOSE: To ensure that all components of a survey system will function as intended when implemented in the full-scale survey.

KEY TERMS: edit, estimation, field test, frame, imputation, instrument, pretest, response rate, stage of data collection, survey, survey system, and variance.

STANDARD 2-4-1: One type of a pretest is a pilot test in which some components of a survey system can be pretested prior to a field test of the survey system (for example, focus groups, cognitive laboratory work, pilot tests, and or calibration studies). New concepts and survey questions should be piloted prior to field testing and/or full-scale implementation. (see Standard 2-1-1, 11)

STANDARD 2-4-2: A second type of pretest is a field test. Components of a survey system that cannot be successfully demonstrated through previous work must be field tested prior to implementation of the full-scale survey. The design of a field test must reflect realistic conditions, including those likely to pose difficulties for the survey. Documentation of the field test (e.g., materials for technical review panels, working papers, technical reports) must include the design of the field test; a description of the procedures followed; analysis of the extent to which the survey components met the pre-established criteria; discussion of other potential problems uncovered during the field test; and recommendations for changes in the design to solve the problems.

GUIDELINE 2-4-2A: Elements to be tested and measured may include alternative approaches to accomplishing a particular task. Elements to be tested may include frame development; sample selection; questionnaire design; change in question order; data collection; response rates; data processing (e.g., entry, editing, imputation); estimation (e.g., weighting, variance computation); file creation; and tabulations.

GUIDELINE 2-4-2B: For an ongoing survey, a change in the question order, the addition of new elements or content, and changes in elements resulting from an evaluation of the survey (see Standard 4-3) should be field tested.

GUIDELINE 2-4-2C: The evaluation criteria for a successful field test should be developed before the field test begins. Key evaluation criteria are established during the design stage. If the criteria are not met, that survey component should not be implemented without field testing a redesigned component.

GUIDELINE 2-4-2D: The results of a field test should be available and analyzed for internal use prior to making a decision to implement the full-scale survey or assessment.

GUIDELINE 2-4-2E: Survey design and instrumentation should be revised to reflect modifications suggested by the results of the field test. A revised budget should be developed, if necessary, to reflect both changes in design and knowledge gained during the field test about resource requirements.

SUBJECT: MAINTAINING DATA SERIES OVER TIME

NCES STANDARD: 2-5

PURPOSE: To maintain and report NCES data series that are consistent over time.

KEY TERMS: bridge study, consistent data series, crosswalk study, key variables, and survey.

STANDARD 2-5-1: NCES must maintain and report on a consistent set of data series that may be analyzed over time. Ongoing data collections must maintain and report on a consistent set of key variables, which are based on consistent data collection procedures.

> **GUIDELINE 2-5-1A:** Identify the basic key variables to be assessed on a regular basis to address policy issues and other information needs.
>
> **GUIDELINE 2-5-1B:** Provide estimates of both change and level for time series data in reports. For survey reports, consider publishing 3 or more years of the time series data along with the current year to highlight the time series.
>
> **GUIDELINE 2-5-1C:** Provide a list of other publications containing the data for previous years in the appendix of a survey report.

STANDARD 2-5-2: Continuous improvement efforts sometimes result in a trade-off between the desire for consistency and a need to improve a data collection. If changes are needed in key variables or survey procedures for data series, a plan must be developed that provides the justification or rationale for the changes in terms of their usefulness for policymakers, conducting analyses, and addressing information needs. The plan must also describe adjustment methods, such as crosswalks and bridge studies that could be used to preserve trend analyses. Studies of adjustment methods must be conducted when changes occur.

SUBJECT: EDUCATIONAL ASSESSMENT AND TESTING

NCES STANDARD: 2-6

PURPOSE: To ensure that educational tests used in NCES surveys for measuring and making inferences about education-related domains are valid, technically sound, and fair. To ensure that the administration and scoring of educational tests are standardized, the scales used over time are stable, and the results are reported in a clear unbiased manner.

KEY TERMS: accommodation, assessment, classical test theory, cut score, derived score, Differential Item Functioning (DIF), disability, domain, equating, fairness, field test, Individualized Education Plan (IEP), instrument, Item Response Theory (IRT), linkage, precision, reliability, scaling, scoring/rating, Section 504, Section 508, survey, test blueprint, validity, and vertical scaling.

STANDARD 2-6-1: Instrument Development—All test instruments used in NCES assessment surveys must be developed following an explicit set of specifications. The development of the instrument must be documented so that it can be replicated. The instrument documentation must include the following:

1. Purpose(s) of the instrument, including the educational framework;
2. Domain or constructs that will be measured and the scores that will be produced;
3. Vertical scaling of tests in a longitudinal study;
4. Framework or blueprint of the instrument in terms of items, tasks, questions, response formats, and modes of responding;
5. Number of items and time required for administration;
6. Context in which the instrument will be used;
7. Characteristics of intended participants;
8. Desired psychometric properties of the items, and the instrument as a whole;
9. Conditions and procedures of administering the instrument;
10. Procedures for item analysis, item reduction, calibration of final instrument, and scoring; and
11. Reporting of the obtained scores.

GUIDELINE 2-6-1A: Relevant experts should review the domain definitions and the instrument specifications. The qualifications of the experts, the process by which the review is conducted, and the results of the review should be documented.

GUIDELINE 2-6-1B: All items should be reviewed before and after pilot and field tests. Pilot and field tests should be conducted on subjects with characteristics similar to intended participants. The sample design for pilot and field tests should be documented.

GUIDELINE 2-6-1C: Field test sample should include an adequate number of cases with the characteristics necessary to determine the psychometric properties of items.

GUIDELINE 2-6-1D: Empirical analysis and the model (e.g., Classical and/or Item Response Theory) used to evaluate the psychometric properties of the items during the item review process should be documented.

GUIDELINE 2-6-1E: When a time limit is set for performance, the extent to which the scores include a speed component and the appropriateness of this component to the defined domain should be documented.

GUIDELINE 2-6-1F: If the conditions of administration are allowed to vary across participants, the variations and rationale for them should be documented.

GUIDELINE 2-6-1G: Directions for test administrations should be described with sufficient clarity for others to replicate.

GUIDELINE 2-6-1H: When a shortened or altered form of an instrument is used, the differences from the original instrument and the implications of those differences for the interpretations of scores should be documented.

STANDARD 2-6-2: Validity—All test instruments used in NCES surveys must meet the purpose(s) stated in the instrument specifications. All intended interpretations and proposed uses of raw scores; scale scores, cut scores, linked scores, and derived scores, including composite scores, sub-scores, score differences, and profiles, must be supported by evidence and theory.

GUIDELINE 2-6-2A: Evidence of validity should be based on analyses of the content, response processes (i.e., the thought processes used to produce an answer), internal structure of the instrument, and/or the relationship of scores to a criterion.

GUIDELINE 2-6-2B: The rationale for each intended use of the test instruments and proposed interpretations of the scores obtained should be explicitly stated.

GUIDELINE 2-6-2C: When judgments occur in the validation process, the selection process for the judges (experts/observers/raters) and the criteria for judgments should be described.

STANDARD 2-6-3: Reliability—The scores obtained by a test instrument must be free from the effects of random variations due to factors such as administration conditions and/or differences between scorers. The reliability of the scores must be adequate for the intended interpretations and uses of the scores.

The reliability must be reported, either as a standard error of measurement or as an appropriate reliability coefficient (e.g., alternate form coefficient, test-retest/stability coefficient, internal consistency coefficient, generalizability coefficient). Methods (including selection of sample, sample sizes, sample characteristics) of quantifying the reliability of both raw and scale scores must be fully described. Scorer reliability, rater to rater, and rater-year reliability must be reported when the scoring process involves judgment.

GUIDELINE 2-6-3A: All relevant sources of measurement errors and summary statistics of the size of the errors from these sources should be reported.

GUIDELINE 2-6-3B: When average scores for participating groups are used, the standard error of measurement of group averages should be reported. Standard error statistics should include components due to sampling examinees, as well as components due to measurement error of the test instrument.

GUIDELINE 2-6-3C: Reliability information on scores for each group should be reported when an instrument is used to measure different groups (e.g., race/ethnicity, gender, age, or special populations).

GUIDELINE 2-6-3D: Reliability information should be reported for each version of a test instrument when original and altered versions of an instrument are used.

GUIDELINE 2-6-3E: Separate reliability analyses should be performed when major variations of the administration procedure are permitted to accommodate disabilities.

STANDARD 2-6-4: Fairness—Test instruments used in NCES surveys must be designed, developed, and administered in ways that treat participants equally and fairly, regardless of differences in personal characteristics such as race, ethnicity, gender, age, socioeconomic status, or disability that are not relevant to the intended uses of the instrument.

GUIDELINE 2-6-4A: Language, symbols, words, phrases, and content that are generally regarded as offensive by members of particular groups should be eliminated, except when judged to be necessary for adequate representation of the domain.

GUIDELINE 2-6-4B: Although differences in the subgroups' performance do not necessarily indicate that a measurement instrument is unfair, differences between groups should be investigated to make sure that they are not caused by construct-irrelevant factors.

GUIDELINE 2-6-4C: When research shows that Differential Item Functioning (DIF) exists, studies should be conducted to detect and eliminate aspects of test design, content, and format that might bias test scores for a particular subgroup.

GUIDELINE 2-6-4D: In testing applications where the level of linguistic or reading ability is not a purpose of the assessment, the linguistic or reading demands of the test instrument should be kept to a minimum.

GUIDELINE 2-6-4E: The testing or assessment process should be carried out so that test takers receive comparable and equitable treatment during all phases of the testing process.

STANDARD 2-6-5: Testing individuals with disabilities or limited English proficiency—Whenever possible, scores derived from test instruments used in NCES surveys must validly, reliably, and fairly reflect the performance of all participants, including individuals with disabilities and individuals of diverse linguistic backgrounds. Although the exact procedures will vary across surveys, appropriate and reasonable accommodations in accordance with applicable federal nondiscrimination laws for special populations must be incorporated. Differences in performance must reflect the construct measured rather than any construct-irrelevant factors such as disabilities and/or language differences.

GUIDELINE 2-6-5A: Permitted accommodations and/or modifications for special populations and the rationale for each accommodation should be documented in the data file and survey methodology report.

GUIDELINE 2-6-5B: The extent to which data gathered with accommodations meet measurement standards of validity and reliability should be documented.

For individuals with disabilities:

GUIDELINE 2-6-5C: Empirical procedures used to review items to ensure fairness, to evaluate whether DIF exists, and to determine accommodations for students/individuals with disabilities should be included in the documentation.

GUIDELINE 2-6-5D: Decisions about accommodations for individuals with disabilities should be made by individuals who are knowledgeable of existing research on the effects of the specific disabilities on test performance.

GUIDELINE 2-6-5E: The participant's Individualized Education Plan (IEP) or Section 504 plan must be consulted prior to making determinations of whether a participant with a disability will participate in the assessment, and what accommodations, if any, are appropriate.

For individuals of diverse linguistic backgrounds:

GUIDELINE 2-6-5F: Empirical procedures used to review items to ensure appropriateness of materials for participants with various backgrounds and characteristics (e.g., nativity, experience in U.S. schools) should be documented to

evaluate whether DIF exists, and to evaluate the linguistic or reading demands to ensure that they are no greater than required.

GUIDELINE 2-6-5G: If an instrument is translated to another language, translation evaluation procedures, and the comparability of the translated instrument to the original version should be documented.

STANDARD 2-6-6: Administration—Administration of all test instruments used in each NCES survey must be standardized. Test administration must follow procedures specified in the test administration manual. The administration manual must include descriptions of the following:

1. Brief statement of the purpose of the survey and the population to be tested;
2. Required qualifications of those administering the instrument;
3. Required identifying information of the participant;
4. Materials, aids, or tools that are required, optional, or prohibited;
5. Allowable instructions to the participants and procedures for timing the testing;
6. Assignment of participants to groups, or special seating arrangements, and preparation of participants as relevant;
7. Allowable accommodations;
8. Desired testing conditions/environment; and
9. Procedures to maintain security of the materials, as applicable, and actions to take when irregularities are observed.

GUIDELINE 2-6-6A: Administration procedures should be field tested. The approved procedures should be described clearly so they can be easily followed.

GUIDELINE 2-6-6B: Survey staff administering the instrument should be trained according to the procedures prescribed in the administration manual.

GUIDELINE 2-6-6C: Modifications or disruptions to the approved procedures should be documented so the impact of such departures can be studied.

GUIDELINE 2-6-6D: Instructions presented to participants should include sufficient detail to allow the participants to respond to the task in the manner intended by the instrument developer.

GUIDELINE 2-6-6E: Samples of administration sites should be monitored to ensure that the instrument is administered as specified.

STANDARD 2-6-7: Scoring and Scaling—Test scoring must be standardized within each survey, and scales must be stable if used over time.

GUIDELINE 2-6-7A: Machine-scoring procedures should be checked for accuracy. The

procedure, and the nature and extent of scoring errors, should be documented.

GUIDELINE 2-6-7B: Hand scoring procedures should be documented, including rules governing scoring decisions, training procedures used to teach the rules to the coding staff, quality monitoring system used, and quantitative measures of the reliability of the resulting ratings. Criteria for evaluating the quality of individual responses should not be changed during the course of the scoring process.

GUIDELINE 2-6-7C: All systematic sources of errors during the scoring process should be corrected and documented.

GUIDELINE 2-6-7D: Consistency among scorers and potential drift over time in scoring/rating should be evaluated and documented.

GUIDELINE 2-6-7E: Meanings, interpretations, limitations, rationales, and processes of establishing the reported scores should be clearly described in the technical report.

GUIDELINE 2-6-7F: Stability of the scale should be monitored and corrected or revised, when necessary, if a scale is maintained over time.

GUIDELINE 2-6-7G: Procedures for scoring—raw scores or scale scores—should be documented. The documentation should also include a description of the populations used for their development.

GUIDELINE 2-6-7H: Procedures for deriving the weights should be described when weights are used to develop the scale scores.

GUIDELINE 2-6-7I: Population norms to which the summary statistics refer should clearly be defined when group performance is summarized using norm scores.

GUIDELINE 2-6-7J: Rationales and procedures for establishing cut scores should be documented when cut scores are established as part of the scale score reporting.

GUIDELINE 2-6-7K: Cut scores should be valid; that is, participants above a cut point should demonstrate a qualitatively greater degree and/or different type of skills/knowledge than those below the cut point.

GUIDELINE 2-6-7L: The method employed in a judgmental standard-setting process should be documented. The documentation should include the following:

1. Selection and qualifications of judges;
2. Nature of the request for their judgments;
3. Training provided to the judges;
4. Feedback of information to judges;
5. Opportunities for judges to confer with one another concerning their judgments; and

6. Methods used to aggregate the judgments and translate them into cut scores.

GUIDELINE 2-6-7M: The judgmental methods used to establish cut scores should meet the following three criteria:

1. The judgmental method should involve peer review and pretesting.

2. The judgments to be provided should not be so cognitively complex that the judges are unable to provide meaningful judgments.

3. The process used to set cut scores should be described in sufficient detail so the process can be replicated.

GUIDELINE 2-6-7N: An estimate of the amount of variability in cut scores must be provided regardless of whether the standard-setting procedure is replicated.

GUIDELINE 2-6-7O: Equating/linking functions should be invariant across sub-populations when equating or linking is used to determine equivalent scores. Supporting evidence for the interchangeability of tests/test forms should be provided.

GUIDELINE 2-6-7P: Detailed technical information (i.e., design of equating or linking studies, standard errors of measurement, statistical methods used, size and relevant characteristics of samples used, and psychometric properties of anchor items) should be provided for the methods by which equating or linking is established.

GUIDELINE 2-6-7Q: Users should be warned that scores are not directly comparable when converted scores from two versions of the test are not strictly equivalent.

STANDARD 2-6-8: Reporting—Test results of the testing should be provided with sufficient detail and contextual information to understand the inferences that can and cannot be made from them.

GUIDELINE 2-6-8A: The analysis of item responses or test scores should be described in detail, including procedures for scaling or equating/linking.

GUIDELINE 2-6-8B: Appropriate interpretations of all reported scores should be provided. The interpretations should describe what the test covers, what the scores mean, and the precision of the scores. The generalizability and limitations of reported scores should also be presented. Potential users should be cautioned against unsupported interpretations; that is, interpretations of scores that have not been investigated, or interpretations of scores inconsistent with available evidence.

GUIDELINE 2-6-8C: Validity and reliability should be reported for the level of aggregation for which the scores are reported when matrix sampling is used. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores indicate that reporting individual scores is meaningful.

STANDARD 2-6-9: Manual—All evidence of compliance with the standards set forth above for each test instrument used in NCES surveys must be compiled in a manual.

> **GUIDELINE 2-6-9A:** Technical documentation should provide technical and psychometric information on a test as well as information on test administration, scoring, and interpretation.

REFERENCES

American Educational Research Association (AERA). (1999). *Standards for Educational and Psychological Testing.* Prepared by the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Washington, DC: AERA.

Code of Fair Testing Practices in Education. (1980). Prepared by the Joint Committee on Testing Practices. Washington, DC.

Educational Testing Service (ETS). (2000). *ETS Standards for Quality and Fairness.* Princeton, NJ: Author and Publisher.

U.S. Department of Education, Office for Civil Rights. (July 6, 2000). *The Use of Tests When Making High-Stakes Decisions for Students: A Resource Guide for Educators and Policymakers.* Washington, DC: Author.

U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. (1992, Reprinted May 1996). *NCES Statistical Standards* (NCES 92-02 1). Washington, DC: U.S. Government Printing Office.